Programación de GPUs con CUDA

Alvaro Cuno

23/01/2010

Agenda

- GPUs
- Cuda
- Cuda + OpenGL

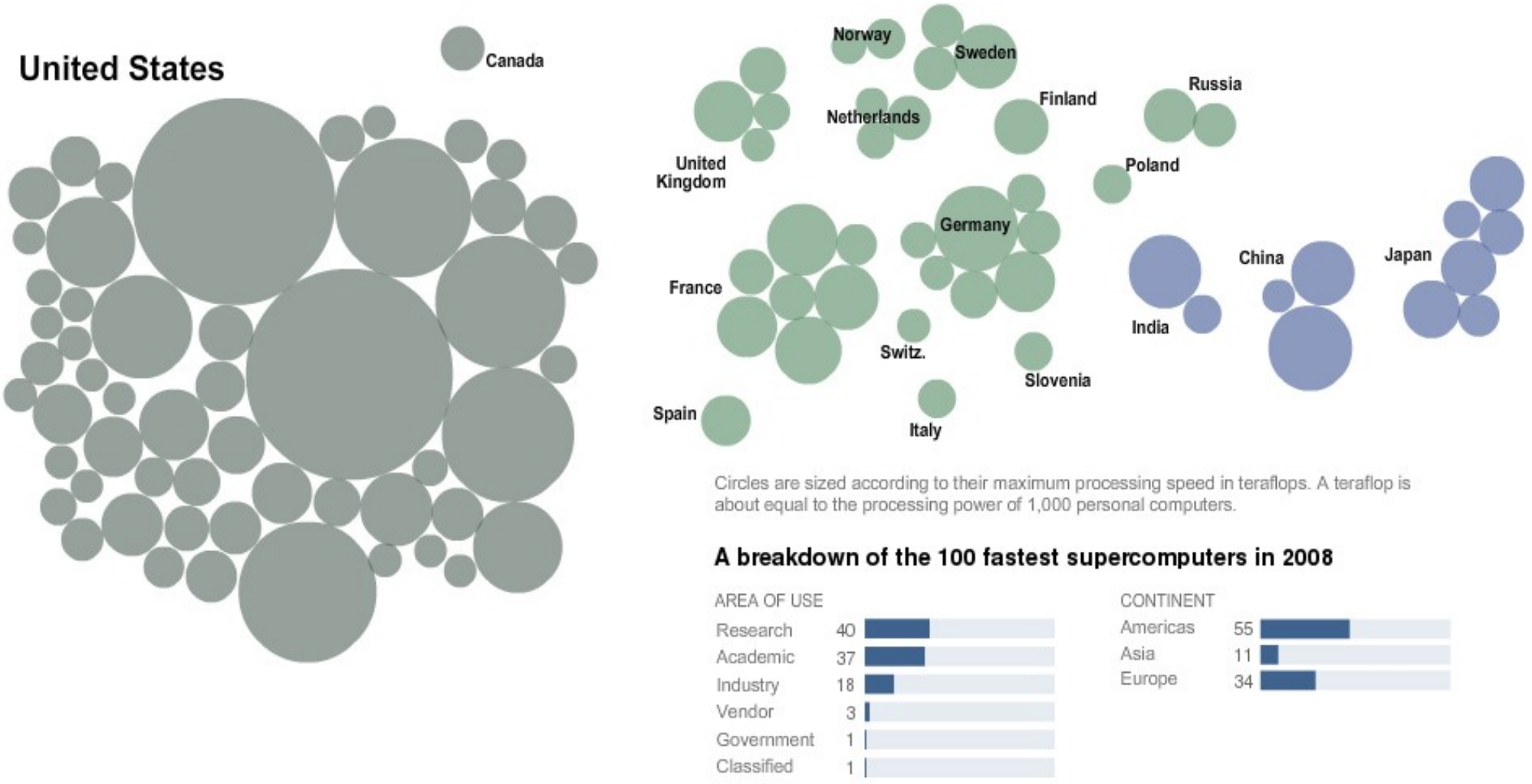
GPUs

(*Graphics Processing Units*)

Supercomputadores

Mapa de los 100 supercomputadores

[1700 teraflops – 27 teraflops] [US$133M -]



Sudamérica: posiciones 306 y 363 (Brasil).

Fuente http://www.top500.org/

Arquitecturas paralelas

- Sistemas con memoria distribuida o compartida
- Clusters
  - Cientos de KiloWatts de consumo
  - Alto costo de instalación y mantenimiento
- Comparado con arquitecturas secuenciales
  - Pocos usuarios (grandes instituciones)
  - Pocas herramientas para desarrollo

Alternativas modernas

- Arquitecturas multi-core/many-core
  - Actualmente 4 cores
  - 128 cores en 12 años
- Arquitectura Cell (Playstation 3)
- Xbox 360 (Microsoft)
- **Placas gráficas (GPUs)**

Computadores con GPUs

- Tesla Computing System
  - 4 GPUs 240 cores c/u
  - 4 TeraFlops SP
  - Precio: US$ 10 000
- Desktop/GPU
  - 1 GPU, 128 cores
  - 470 GigaFlops
  - Precio: US$ 300

Computadores con GPUs

- Tesla Computing System
  - 4 GPUs 240 cores c/u
  - 4 TeraFlops SP
  - Precio: US$ 10 000
- Desktop/GPU
  - 1 GPU, 128 cores
  - 470 GigaFlops
  - Precio: US$ 300

Procesamiento paralelo para las masas!!!

- The 29th most powerful supercomputer is GPU-based technology (170 teraflops)

GPUs

- Evolución producto del insaciable mercado de gráficos 3D de alta calidad (juegos, industria del cine, etc.)

GPUs

- Evolución producto del insaciable mercado de gráficos 3D de alta calidad (juegos, industria del cine, etc.)

1987

2009

GPUs

- Evolución producto del insaciable mercado de gráficos 3D de alta calidad (juegos, industria del cine, etc.)

GPUs

- Evolución producto del insaciable mercado de gráficos 3D de alta calidad (juegos, industria del cine, etc.)

GPUs

- Evolución producto del insaciable mercado de gráficos 3D de alta calidad (juegos, industria del cine, etc.)

2D animation

GPUs

- Evolución producto del insaciable mercado de gráficos 3D de alta calidad (juegos, industria del cine, etc.)
  - Placas gráficas de funcionalidad fija
  - GPUs programables
    - Altamente paralelas
    - Multicores (soporte a múltiples hilos)
    - Altísimo poder de calculo
    - Gran ancho de banda: GPU y su memoria

GPU

- Es un procesador extremamente potente, flexible y de bajo costo

Placa de video Nvidia GeForce 8800GT

GPU

- Es un procesador extremamente potente, flexible y de bajo costo
- Originalmente diseñado para procesamiento gráfico en 3D
  - Rendering of polygons
  - Texturing
  - Shading

Placa de video Nvidia GeForce 8800GT

GPU: poder de cómputo

Intel Core2 Quad 3.0GHz

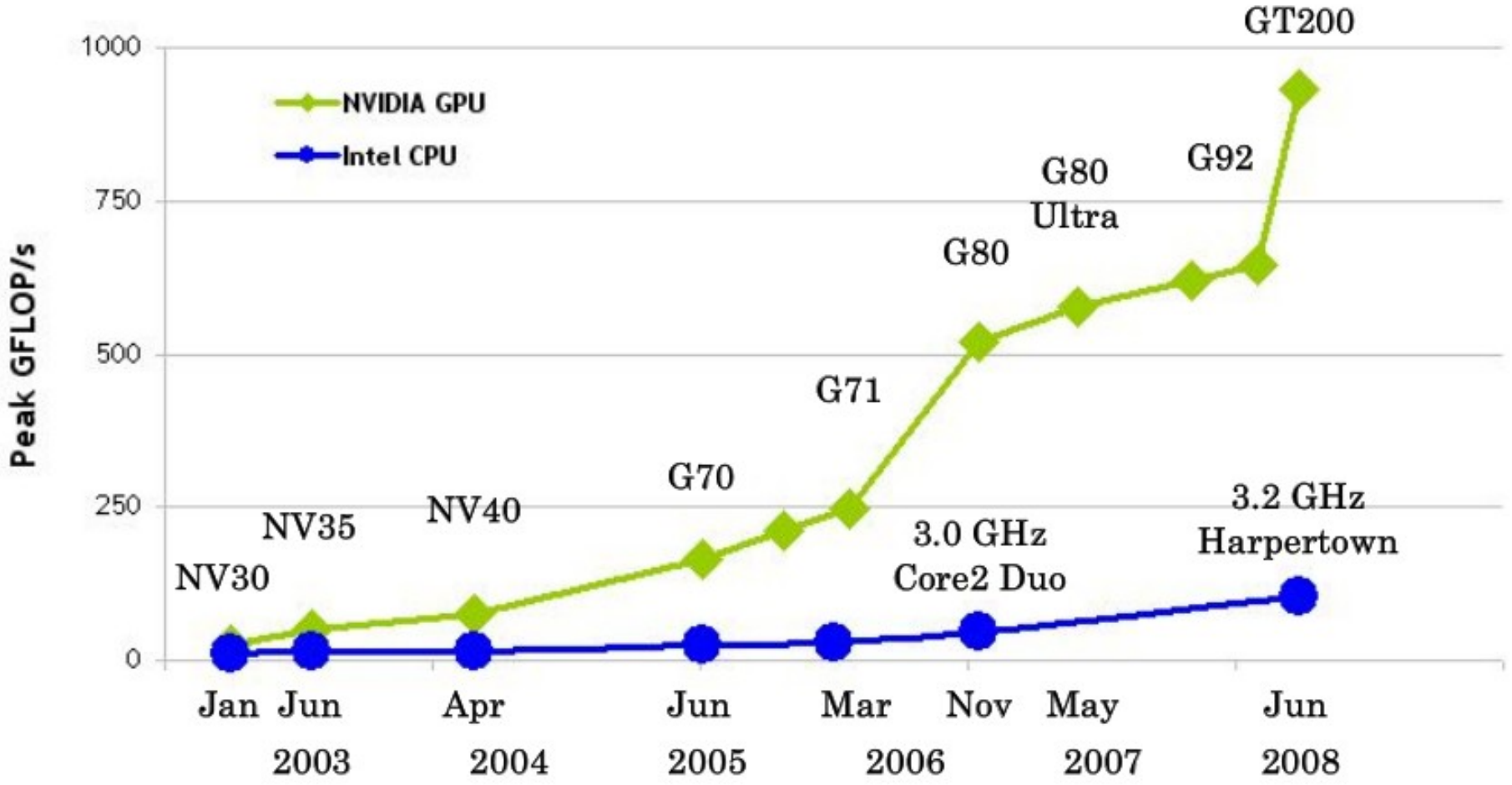
- Computation: 96 GFLOPS
- Memory bandwidth: 21GB/s
- Price: US$ 570*
- Anual growth: 1.4x

NVidia GeForce 8800 GTX

- Computation: 330 GFLOPS
- Memory bandwidth 86.4GB/s
- Price: $300*
- 1.7x (fragment shader), 2.3x (vertex shader)

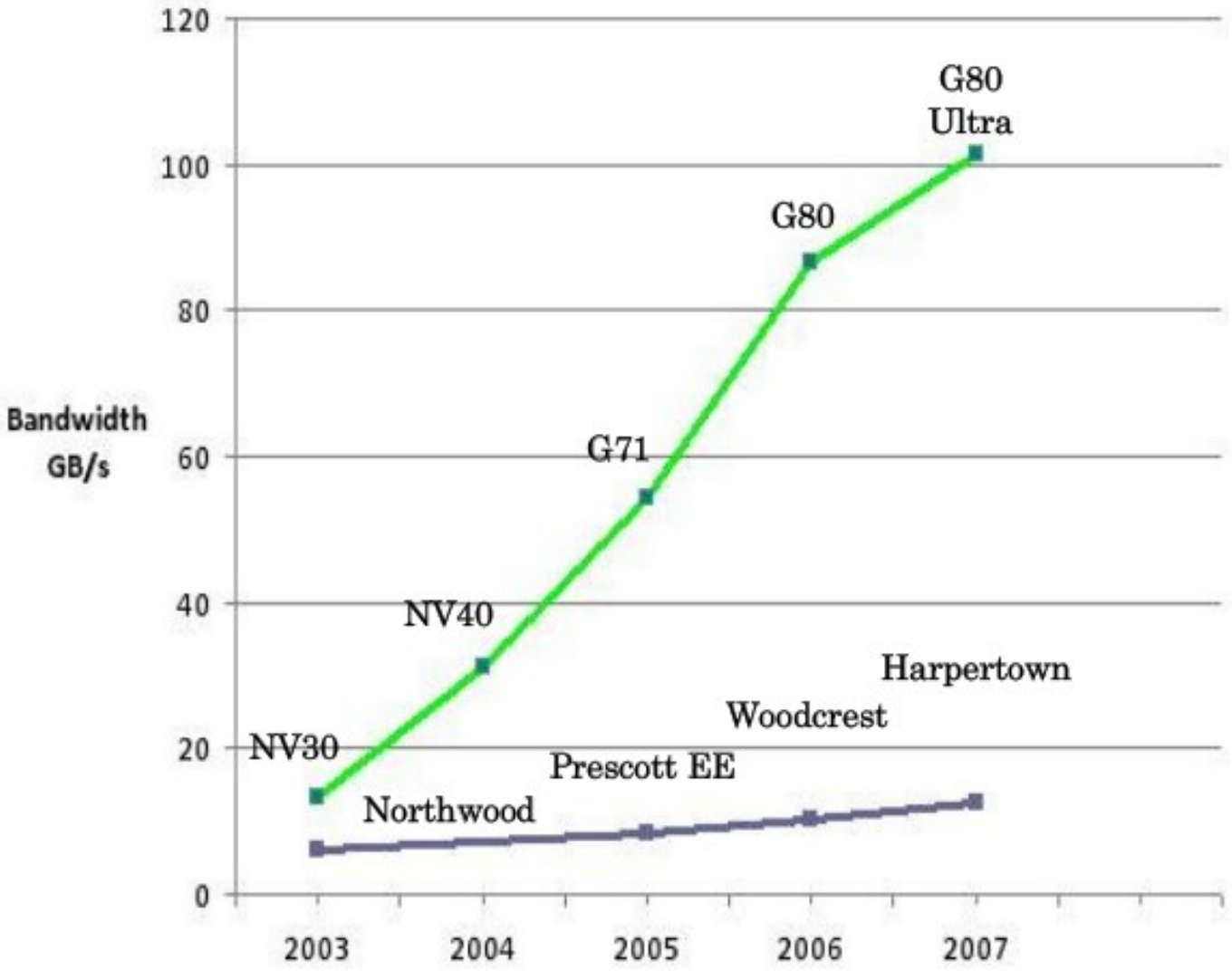
* Precios referidos al año 2008

GPU: poder de cómputo



Fuente: Nvidia CUDA programmer guide

GPU: ancho de banda



Fuente: Nvidia CUDA programmer guide

GPU: flexibilidad

- GPUs son altamente programables
  - Vertex shader, geometry shader y fragment shader
  - Lenguajes de programación de alto nivel

GPU: flexibilidad

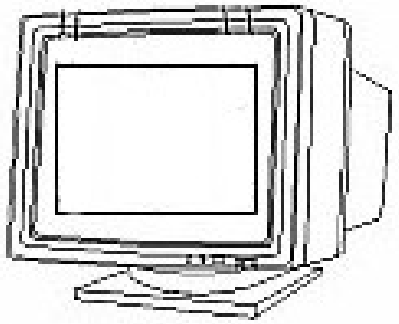
- GPUs son altamente programables
  - Vertex shader, geometry shader y fragment shader
  - Lenguajes de programación de alto nivel
- GPUs recientes soportan cálculos de alta precisión
  - 32 bits en todo el pipeline
  - Precisión suficiente para la mayoría de aplicaciones

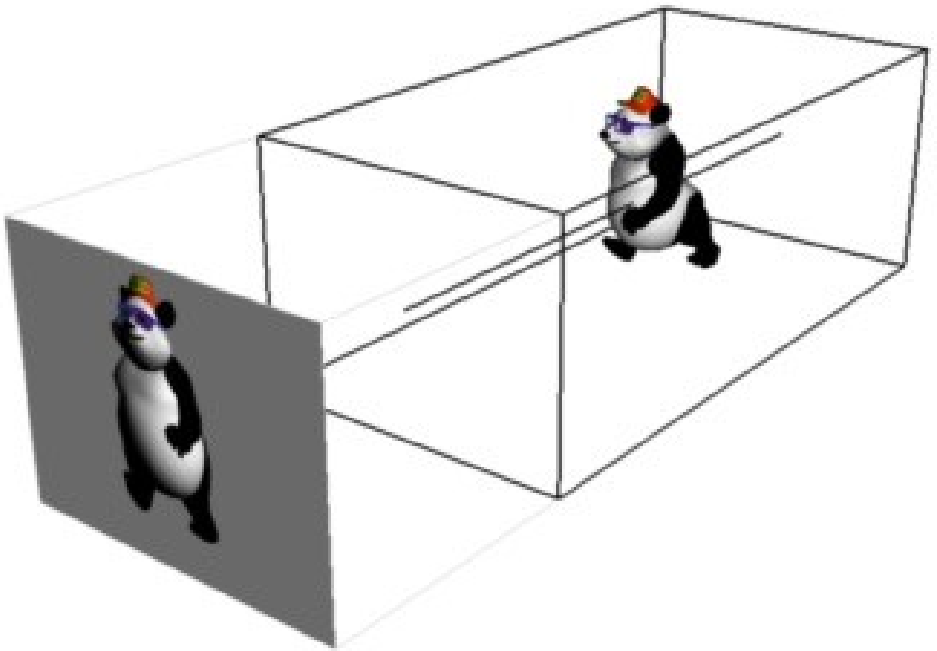
GPU: flexibilidad

- GPUs son altamente programables
  - Vertex shader, geometry shader y fragment shader
  - Lenguajes de programación de alto nivel
- GPUs recientes soportan cálculos de alta precisión
  - 32 bits en todo el pipeline
  - Precisión suficiente para la mayoría de aplicaciones
- Tienen muchos núcleos y una arquitectura mas simple que una CPU estándar

¿Por qué son tan potentes?

- Originalmente diseñados para cálculo matemático y procesamiento paralelo intenso



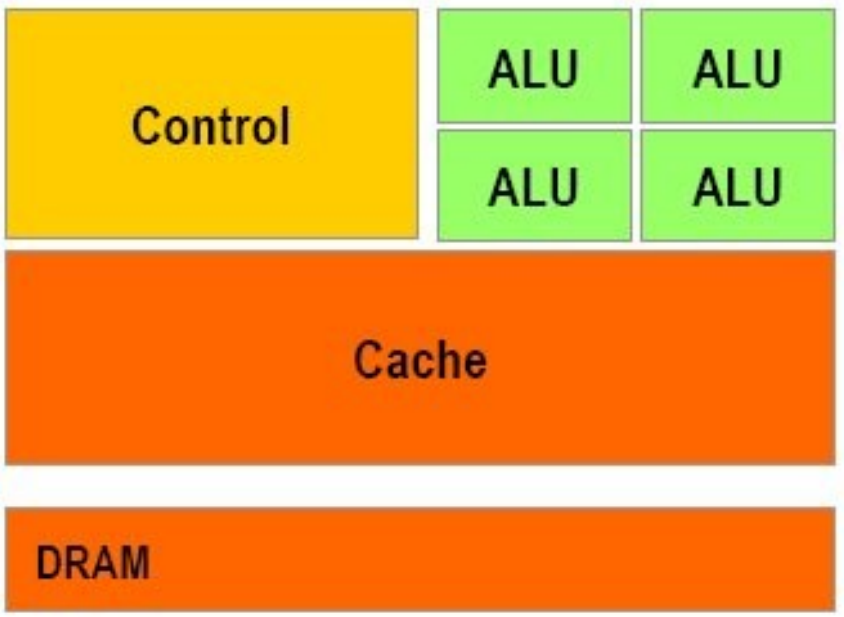


¿Por qué son tan potentes?

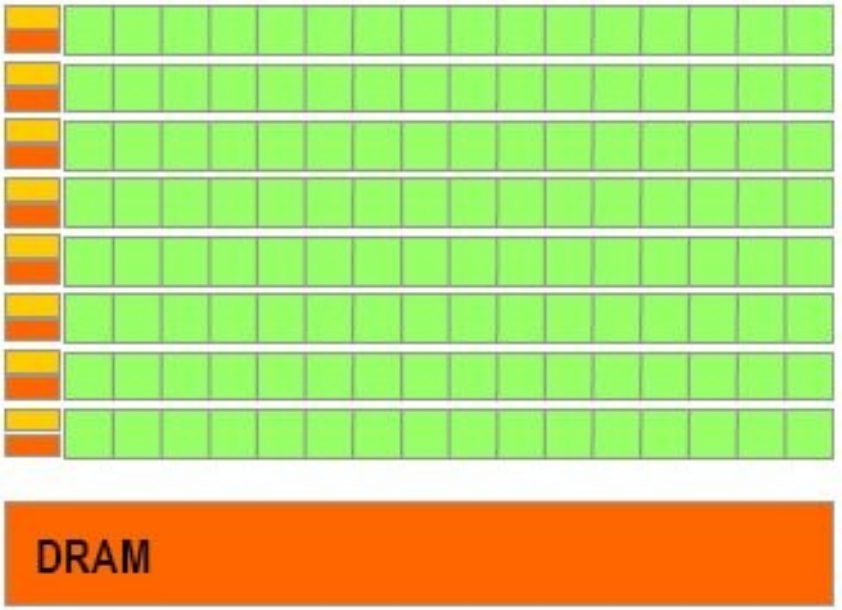
- Originalmente diseñados para cálculo matemático y procesamiento paralelo intenso
- Más transistores para procesamiento que para data caching y control de flujo

¿Por qué son tan potentes?

- Originalmente diseñados para cálculo matemático y procesamiento paralelo intenso
- Más transistores para procesamiento que para data caching y control de flujo



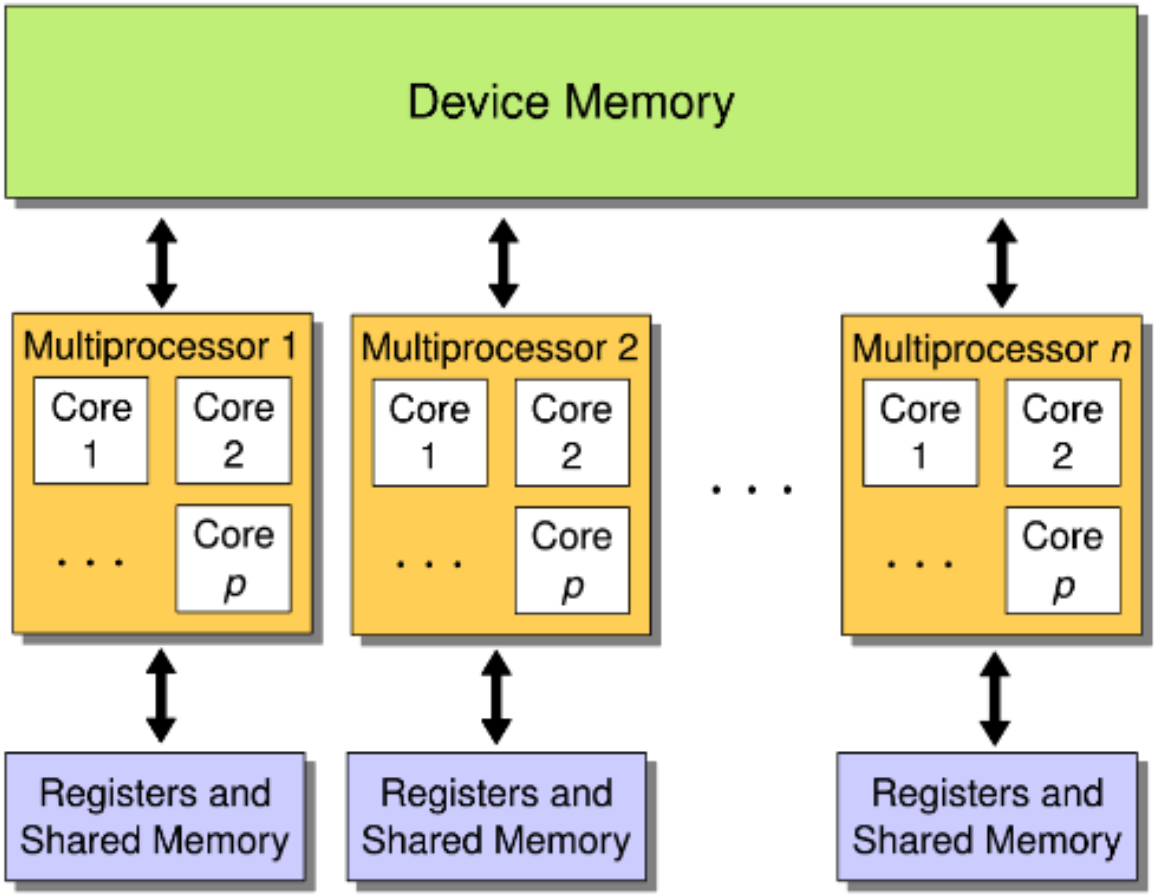
CPU



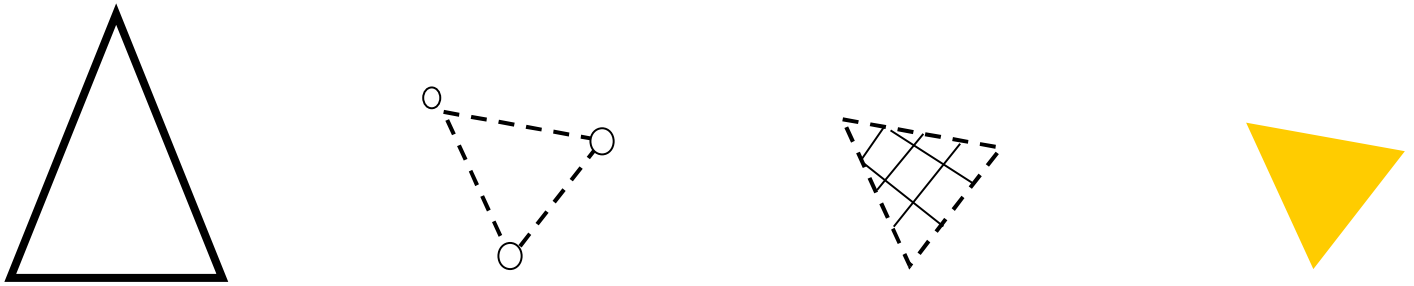
GPU

¿Por qué son tan potentes?

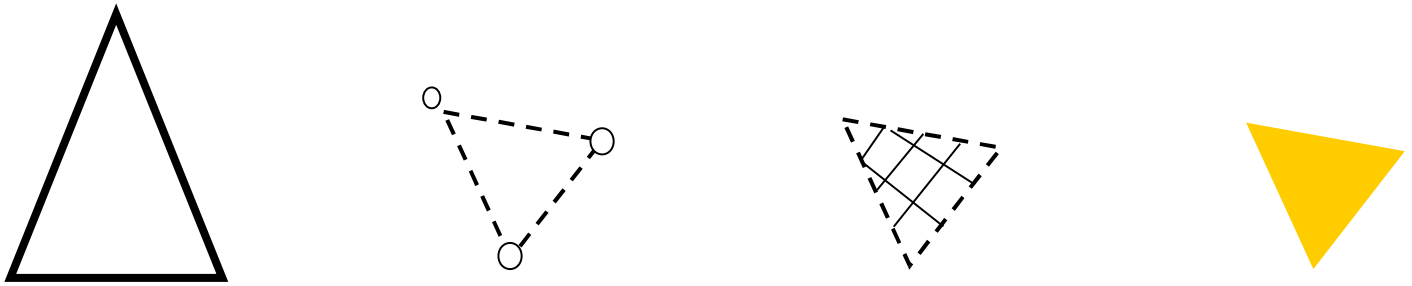
- Arquitectura simple

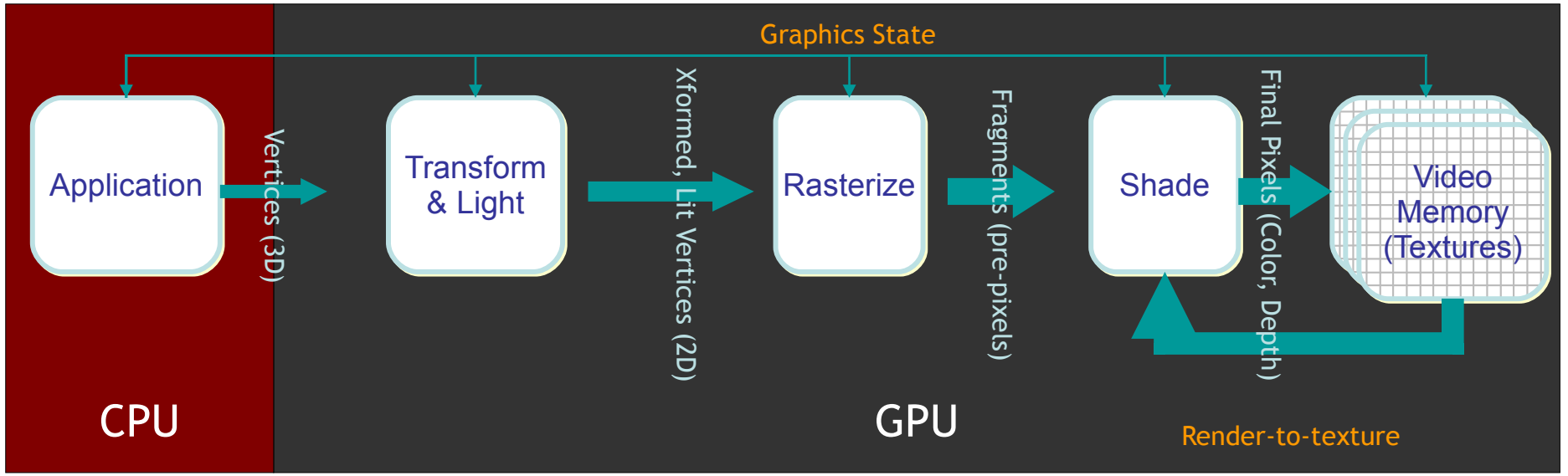


GPU: procesamiento gráfico

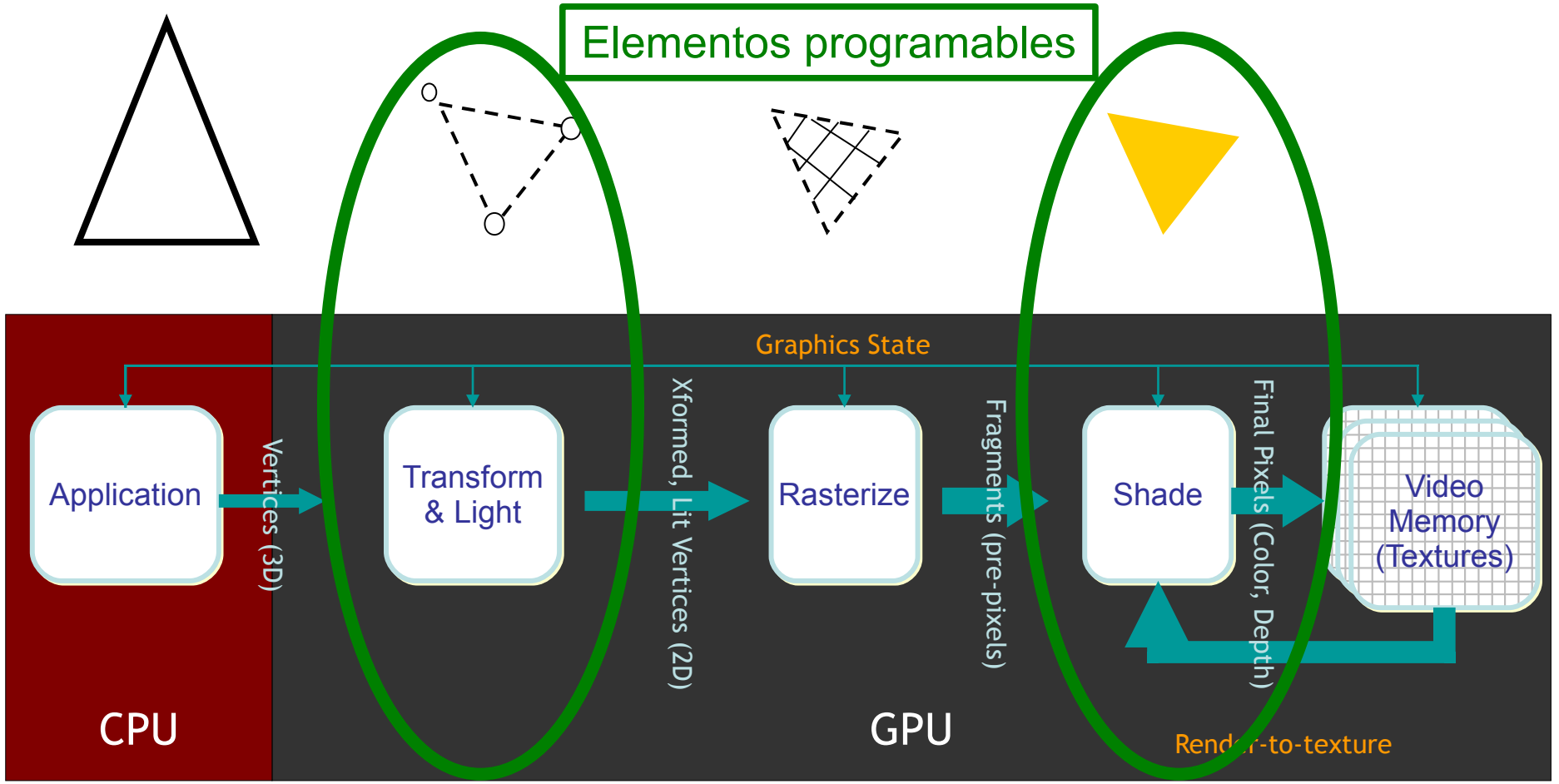


GPU: procesamiento gráfico

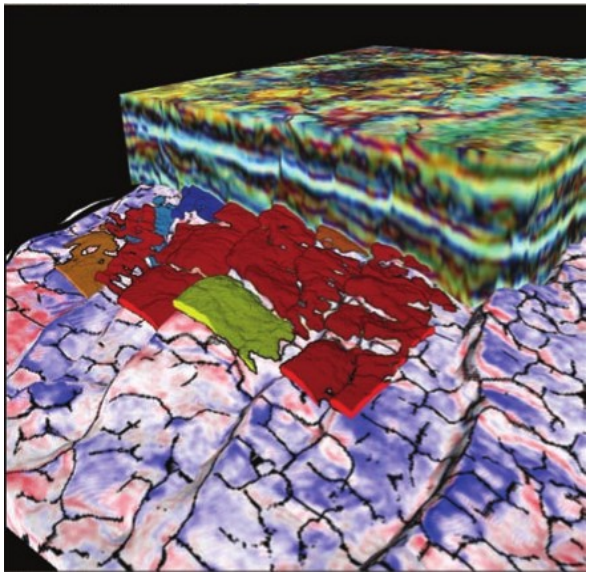


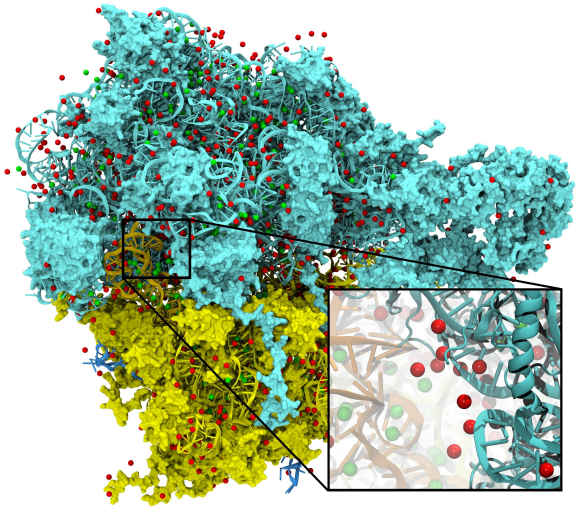


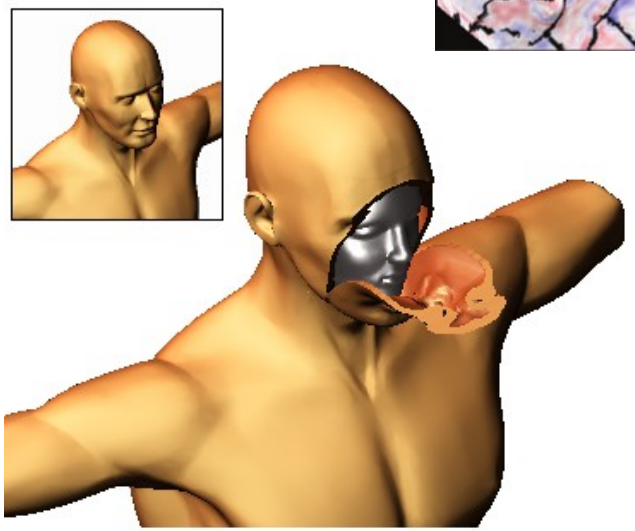
GPU: procesamiento gráfico

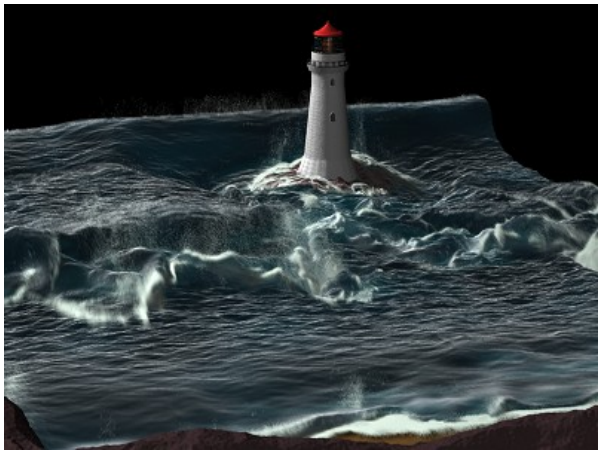


Aplicaciones recientes









Aplicaciones recientes

Aplicaciones recientes

Evolución de las GPUs

Pre-GPU	Soluciones costosas de S/H (SGI)	
1ra Gen	NVIDIA TNT2, ATI Rage, 3DFX Voodoo3	rasterization, no transformations
2da Gen	NVIDIA GeForce256 and GeForce 2, and the ATIRadeon 7500 (1999-2000)	transformation and lighting
3ra Gen	NVIDIA GeForce3 and GeForce4, and the ATI Radeon 8500 (2001-2002)	Vertex shader, limited instructions number, absence of program flow control, fixed-point numbers
4ta Gen	FX series and the ATI Radeon 9700 and 9800 (2003-2006)	Vertex and fragment shaders, texture data, floating-point numbers
5ta Gen	NVIDIA GeForce 8 series (2007-)	Graphics hardware for general-purpose computation

Programación de GPUs

- No es muy simple
- No se trata de ”recompilar” código tradicional
- GPUs fueron diseñadas para procesamiento gráfico
  - Modelo de programación inusual
  - Programación dependiente del pipeline gráfico
    - Cg, GLSL
  - Ambiente de programación restringido
    - CTM(AMD), CUDA(NVIDIA), OpenCL, etc.

Programación de GPUs

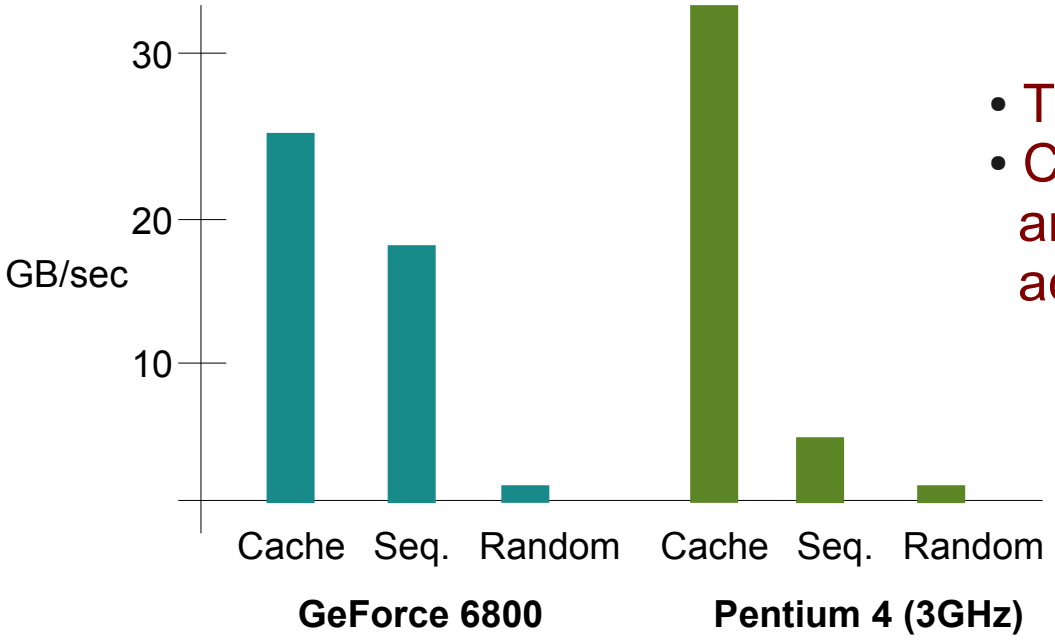
- Arquitecturas de GPUs
  - Ejecutan procesamiento paralelo
  - Evolución constante
  - Secreto del fabricante

Características de las GPUs

- Modelos de acceso a memoria
- Intensidad aritmética
- Transferencia de datos
- Operaciones típicas

Características de las GPUs

- Modelos de acceso a memoria
  - Cached, Sequential, Random



- Then … locality, locality, locality
- Computation must be structured around sequential memory accesses

Comparación de desempeño de memoria: GPU vs. CPU

Características de las GPUs

- Intensidad aritmética
  - GPU fragment processor = 60 Gflops
  - Pentium4 = 12 Gflops
  - More computation! ... how much?
  - To cover memory latency our programs need to contain enough arithmetic instructions
  - AI = arithmetic_op/memory_op
  - Write algorithms with high AI !

Características de las GPUs

- Transferencia de datos entre la CPU y GPU
  - PCI: 3.2 GB/sec
  - Vector addition: A+B = C
  - How to avoid this penalty?
  - Amortize the data transfer cost

Características de las GPUs

- Operaciones típicas
  - Map
    - M’ = kM, simple to implement on the GPU
  - Reduce
    - N to 1, multiple steps
  - Stream filtering
    - Non-uniform reduction, involves removing items
  - Sort/Search
  - Gatter(get) / Scatter(set)

Características de las GPUs

- Algoritmos paralelizables: aceleración > x10
  - Algoritmos que aplican la misma función a una gran cantidad de datos
  - Algoritmos que requieren cálculos de 32 bits
    - 64 bits es soportado
  - Algoritmos donde el monto de comunicación entre la CPU y la GPU es mínimo
- No todo algoritmo es paralelizable!